Segmentation Data Analysis

Jeffrey D. Scargle

Space Science Division, NASA Ames Research Center

MS 245-3

Moffett Field, CA, 94035-1000

Abstract- To extract information from measurements distributed over a data space it is useful to find an optimal way to partition the measurements and fit a simple model to each of the partition elements. This paper presents an algorithm for finding such optimal segmentation models, with examples from several fields.

I. Introduction: The Data

Much data in the fields of astrophysics, Earth and space science are in the form of measurements distributed over a data space of known dimension. The following examples underscore the great variety of such situations:

- time series data – 1D
- spectra -- 1D
- image data -- 2D
- redshift galaxy surveys – 3D
- X- and γ-ray photon data -- 4D+

In the last case the dimensions are space, time, energy, and possibly additional data. The measurements may refer to a physical quantity measured over a predefined sub-interval of the data space. The sub-interval can be predefined, as in the case of pixels or bins, or defined by the data themselves (c.f. the Voronoi cell examples below). Or one may have point data, where the recorded information comprises the actual locations of the points in the data space – typically, in the form of positional coordinates. An example of this case is a map of high-energy radiation, in the form of positions of photons on the sky. We will see that one way to deal with such data modes is to construct analogs of pixels in the form of Voronoi cells, much like data-defined bins containing one event or point.

II. The Density Function

For the class of data modes outlined above, science analysis can almost always be based on a density estimation procedure. The term density is to be taken in a general sense, including radiation intensity (in space or as a function of wavelength), number of objects per unit volume, or any other quantity expressing signal strength per unit volume of data space. The original density estimation is typically followed by a process to extract quantitative (local or global) information from the density map. For example, one may want to identify specific localized sources, clusters of sources, or other more global structures.

III. The 1D Algorithm

A optimal segmentation algorithm, new but with a dynamic programming heritage dating back to Richard Bellman in the 1950's, can be used to detect structure on any scale in sequential data. When the model fitness function to be optimized is the marginal posterior of the piece-wise constant model, the method (called *Bayesian Blocks*) has been applied to a number of problems in 1D

astronomical data analysis. The figure below depicts the Bayesian block representation of a time series from a gamma-ray burst observed by the NASA Swift mission. The raw data consist of arrival times of individual photons, too numerous to plot individually. The blocks indicated by horizontal solid lines form the optimal step-function model for the photon – i.e., that which maximizes the posterior probability over all models consisting of constant Poisson rates in blocks. The number of blocks is determined by the automatic "Occam factor" of Bayesian analysis. It is mediated by a prior distribution for the number of blocks, but is not based on an *ad hoc* complexity penalty, as in other methods.

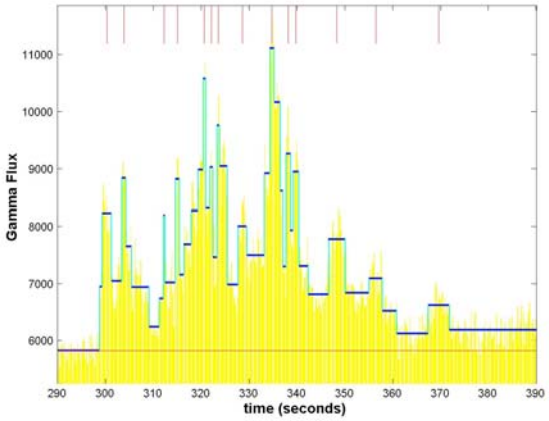

Figure 1: Bayesian block representation of a gamma-ray burst time profile.

Note that the background level is not modeled separately, but is indicated by the first block, presumed to cover the time before the burst actually started.

We have recently explored several other cost functions (Scargle, Norris and Jackson, in preparation); a simple maximum likelihood quantity that has a useful scaling property.

IV. The 2D+ Algorithm

For cost functions that have a simple convexity property, extension to higher dimensional data spaces is immediate (Jackson and Scargle, 2007). Begin with the Voronoi tessellation of the data points, and order the cells by area (or by volume in higher dimensional problems). This one dimensional array is passed to the algorithm described above. The resulting 1D blocks may contain several disconnected fragments in the original higher dimensional data space, but it is straightforward to identify such block fragments and thereby construct a set of simply connected blocks. The job is not then complete, but requires assembly of the blocks in to scientifically meaningful structures. In most cases this step requires the automated invocation of domain knowledge (e.g., point sources must be circularly symmetric and otherwise consistent with the known point spread function of the instrument).

V. Applications

The optimal segmentation algorithm described above, yielding the best-fitting piecewise constant model of the data, has been applied in problems of various dimensions. Examples of such analysis include:

1) Light-curves of gamma-ray bursts (as in the figure above).
2) Cluster and other structural analysis of the large scale distribution of galaxies with the Sloan Digital Sky Survey
3) Point source identification and characterization in gamma ray data (for GLAST, the Gamma Ray Large Area Space Telescope, to be launched in November, 2007)
4) Anomaly detection for a homeland security problem (detectiona of anomalous events in domestic water distribution systems)

Case 1) was depicted in Figure 1 above. Figure 2 is a blocky estimate of the density of galaxies

in the Universe, using data from the Sloan Digital Sky Survey redshift catalog. Of course the full aspect of the 3D distribution can't be seen here, but one does get an impression of a complex web of connected structures of various shapes and sizes – clusters, strings, sheets and voids.

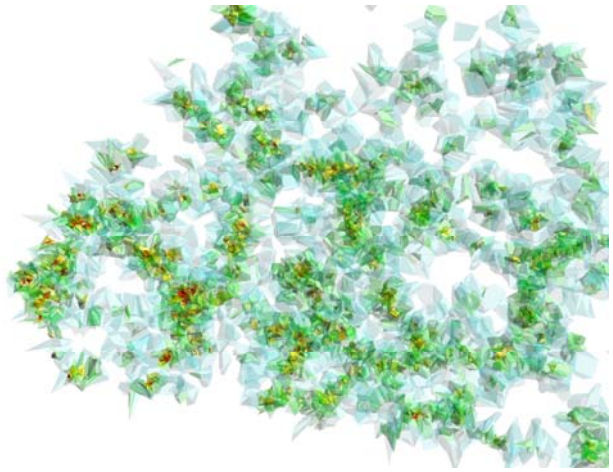

Figure 2: Bayesian block representation of the large-scale structure in the galaxy distribution, based on a portion of SDSS redshift data.

Detection of point sources in gamma-ray survey data such as will be produced by the GLAST mission is complicated by various factors: spatially the photon coordinates lie on a sphere, not a flat plane; each photon has a different point-spread function depending on its energy; point and extended sources, overlapping each other, need to be disentangled; transient and rapidly variable sources need to be detected and studied; and more.

Figure 3 is the first step in a method for detecting sources, both constant and variable. It is a tessellation of the sky into triangles, closely related and informationally equivalent to the Voronoi tessellation. Except for representing segments of great circles as straight lines, this construction is exact. It is based on embedding the 2D points in a 3D spherical surface and computing the convex hull of the resulting configuration.

The diagonal array of small Delaunay triangles occurs in the galactic plane, where sources are more numerous.

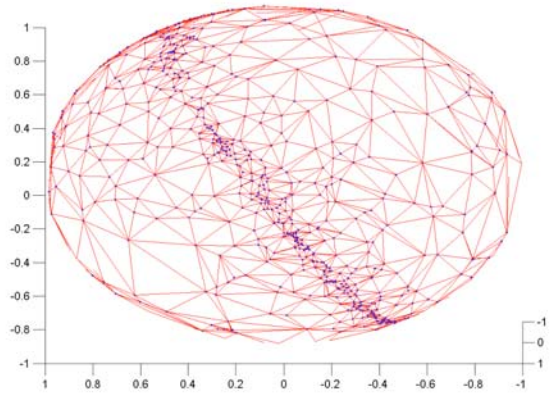

Figure 3: The Delaunay triangulation of synthetic GLAST photon data on the celestial sphere. This construction contains the same information (local density and its gradient, and adjacency information among the photons) as does the Voronoi tessellation, and is easily calculated on the sphere.

Point and extended sources can be identified by a 2D optimal segmentation analysis of this set of photon cells, much as in Fig. 2. But to detect and analyze variability we need to introduce time as a third dimension. A convenient way to do this is shown in Fig. 4: time is represented as a radial coordinate. The start time and end time of the observations lie on a unit sphere and a sphere of some larger radius. Constant sources can be seen by eye in this display, in the form of clusters of points lying closely on radii between the spheres. Transient sources lie on subsegments of the radii, and variable sources have a lumpy distribution on the radii corresponding to their positions on the celestial sphere.

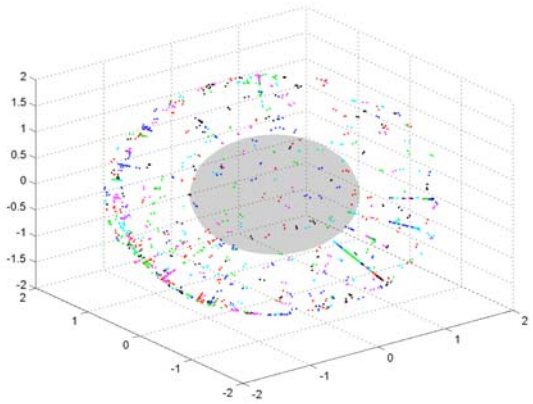

Figure 4: A spherical representation of time and location on the celestial sphere.

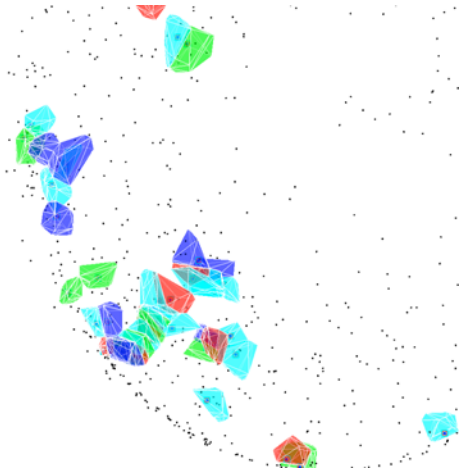

Figure 5: Voronoi tessellation of 3D (space-time) coordinates of synthetic gamma-ray photons. Only a few of the Voronoi cells and the points defining them are plotted for ease of visualization.

Then one can perform a 3D Voronoi tessellation of the photons, as depicted in Fig. 5. Block analysis of these cells then can be used to detect and characterize variable sources. For example, Fig. 6 shows a dynamic source detection result for synthesis of 55 days of GLAST data. Early in the mission, only the brightest sources have been detected, but as time goes along, more and more sources achieve a statistically significant detection, and variable sources can be characterized.

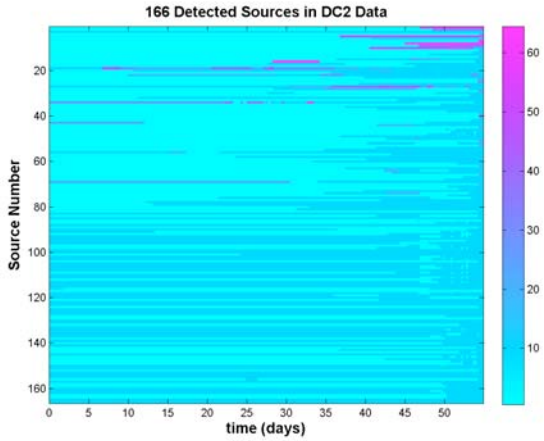

Figure 6: A synthetic real-time list of sources that might be detected in 55 days of the GLAST mission. The color scale represents the number of photons in the most intense block of the (point) sources. The vertical axis is an arbitrary source identification index.

I conclude with a somewhat different example, namely a way to detect anomalous behavior in multidimensional data. The underlying principle is simply that anomalous points by defintion lie in a region of the data space that has not been populated very much by the previous "normal" data. Hence the Voronoi cells of the anomalous points will be larger on average than those of normal points. One simply tessellates the normal or training data, establishes a threshold of Voronoi cell volume, and deems a new point with a cell volume that exceeds threshold as anomalous. This approach was developed as part of a project to monitor multivariate water quality data for possible intrusive events.

Acknowledgements: I am grateful to Jay Norris and Brad Jackson for various contributions, and to members of the Gamma Ray Large Area Space Telescope (GLAST) project for encouragement and simulated data.

References:

Jackson, Brad; Scargle, Jeffrey D.; Barnes, David; Arabhi, Sundararajan; Alt, Alina; Gioumousis, Peter; Gwin, Elyus; Sangtrakulcharoen, Paungkaew; Tan, Linda;

Tsai, Tun Tao, An Algorithm for Optimal Partitioning of Data on an Interval, IEEE Signal Processing Letters, (2005), **12**, 105-108.

Scargle, J. D., Norris, J., and Jackson, B, (2007), "Studies in Astronomical Time Series Analysis. VI. Optimum Partition of the Interval: Bayesian Blocks, Histograms, and Triggers", (in preparation)

Jackson, B, and Scargle, B, “Optimal Partitioning in Higher Dimensions,” (2007), in preparation.